View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PhilSci Archive

Group Agency and Artificial Intelligence

Christian List*

Discussion paper / Spring 2019

The aim of this exploratory paper is to discuss a sometimes recognized but still under-appreciated parallel between group agency and artificial intelligence. As both phenomena involve non-human goal-directed agents that can make a difference to the social world, they raise some similar moral and regulatory challenges, which require us to rethink some of our anthropocentric moral assumptions. Are humans always responsible for those entities' actions, or could the entities bear responsibility themselves? Could the entities engage in normative reasoning? Could they even have rights and a moral status? I will tentatively defend the (increasingly widely held) view that, under certain conditions, artificial intelligent systems, like corporate entities, might qualify as responsible moral agents and as holders of limited rights and legal personhood. I will further suggest that regulators should permit the use of autonomous artificial systems in high-stakes settings only if they are engineered to function as moral (not just intentional) agents and/or there is some liability-transfer arrangement in place. I will finally raise the possibility that if artificial systems ever became phenomenally conscious, there might be a case for extending a stronger moral status to them, but argue that, as of now, this remains very hypothetical.

1. Introduction

Group agency and artificial intelligence are two much discussed phenomena. The first consists in the fact that certain organized collectives, such as corporations, courts, and states, can function as intentional, goal-directed agents in their own right, over and above their individual members, often with considerable power and influence in the world.[1] The second consists in the fact that computational or robotic systems increasingly display autonomous behaviours and cognitive and practical capacities similar in some respects to those traditionally associated with humans.[2] Examples range from driverless cars and self-navigating drones to autonomous systems in medical, financial, military, and other high-stakes settings. But while group agency and artificial intelligence have each received much attention, the two phenomena are less often considered together. As a theorist of group agency, I see a salient parallel between them, and the aim of this exploratory paper is to discuss this parallel and to draw attention to some lessons for ethics and artificial intelligence that can be learnt from it. In particular, the established work on group agency offers a useful template for thinking through some of the moral and regulatory challenges raised by artificial intelligence. The parallel has been recognized before, notably in

* This work was presented at the workshop on "Artificial Agency and Collective Intelligence", Trinity Hall, Cambridge, September 2017, at the Political Theory workshop at the Forschungskolleg Humanwissenschaften, Bad Homburg, February 2018, and as the 9th Munich Lecture in Business Ethics, February 2019. I am grateful to the participants for helpful comments and feedback.

[1] See, e.g., French (1984), Rovane (1997), Pettit (2001, ch. 5, 2003), Tollefsen (2002, 2015), List and Pettit (2006, 2011), and Tuomela (2013). My discussion of group agency and group responsibility also draws on List (2019).

[2] For overviews, see Russell and Norvig (2009), Winfield (2012), and Boden (2016).

philosophical work by Migle Laukyte and in legal work by Lawrence Solum, and there is also much interest among artificial-intelligence researchers in the theme of multi-agent systems. But the parallel nonetheless warrants further discussion.[3]

In brief, the parallel lies in the fact that group agency and artificial intelligence each involve entities distinct from individual human beings that qualify as intentional agents, capable of acting more or less autonomously in pursuit of certain goals and making a difference to the social world. If such non-human agents can make high-stakes decisions and perform actions on their own, we must answer some salient moral and regulatory questions, which are surprisingly similar in the corporate and artificial cases. Are corporate and artificial entities proper targets for the attribution of responsibility? Could they be *moral* agents, not just *intentional* ones? Could they engage in normative reasoning? Could they attain some kind of legal personhood status and have certain rights? Might we even have to care about them?

In Sections 2 and 3, I will introduce group agency and artificial intelligence in more detail. In Section 4, I will explain the parallel from a conceptual perspective and suggest that group agency can be interpreted as a special case of artificial intelligence. In Sections 5 and 6, I will defend the view that, under certain conditions, artificial entities, like corporate agents, may qualify as bearers of responsibility and as holders of limited rights and legal personhood, and I will suggest that regulators should demand that when autonomous artificial systems are used in high-stakes settings, they should be engineered to achieve moral – not just intentional – agency. As a backup, there should be liability-transfer arrangements in place. I will further raise the possibility that if we ever encountered phenomenally conscious AI systems, they might become candidates for a stronger moral status, but emphasize that this scenario remains very hypothetical. In Section 7, finally, I will look at the bigger picture for ethics. One lesson is that if we take the arrival of novel, non-human agents seriously, we may have to adjust some of our conventional anthropocentric approaches to morality and regulation so as to accommodate new *loci* of moral agency, responsibility, and even rights.

---

[3] The most extensive prior discussion of this parallel that I am aware of can be found in Laukyte (2014, 2017), where themes similar to the present ones are addressed, drawing on the theory of group agency in List and Pettit (2011). That said, I developed my account of the parallel independently, and the motivating example of an agent in the List-Pettit book itself is a robot. The parallel was also implicit in the themes of the 2017 Cambridge workshop at which the present work was first presented (see the acknowledgment note) and was one of the motivations for an LSE workshop on "The Politics and Philosophy of Artificial Intelligence" in June 2014 that I co-organized with Kai Spiekermann. An earlier discussion of the parallel from a legal perspective can be found in the work of Lawrence Solum on AI legal personhood, which is often modelled on the idea of corporate legal personhood. See especially Solum (1992), as well as follow-up contributions. There is sizeable literature in computer science and artificial intelligence on multi-agent systems, though that literature typically focuses more on the interplay between multiple agents and less on the question of the emergence of intentional agency at the level of the multi-agent system as a whole. See, among others, Chopra, van der Torre, and Verhagen (2018).

2. Group agency

Group agency is the phenomenon that suitably organized collectives can constitute intentional, goal-directed agents in their own right, over and above their individual members.[4] Examples are firms and corporations, collegial courts, universities, churches, NGOs, and even states in their entirety. We tend to ascribe beliefs, desires, and other intentional attitudes to such entities, and treat them as *loci* of agency, sometimes even as artificial persons with obligations and a legal status, as distinct from the obligations and status of the people who belong to them.

At first, one might think that the ascription of agency and intentional attitudes to certain collectives is just a metaphor. Saying that Microsoft intends to increase profits is a shorthand for saying that the managers, board members, or shareholders have this intention. Likewise, saying that the Roman Catholic Church is committed to certain values is a shorthand for saying that individual clergymen and other Catholics have those commitments. But there is a fairly straightforward indispensability argument for treating group agency as not just metaphorical but real. According to this argument, a realist view about group agency is supported by the way we speak about many collectives in ordinary as well as social-scientific discourse. Our best social-scientific theories represent some collective entities as goal-directed agents and explain their behaviour by using the same concepts and categories that we use to explain individual behaviour. For example, the theory of the firm in economics and "realist" theories of international relations apply standard rational-actor models to firms and states. In fact, a profit-maximizing firm may be a more fitting case of a *homo economicus*, a self-interested utility-maximizing rational agent, than any individual human being is. And strategic interactions between states, such as between the United States and the Soviet Union during the Cold War, are often modelled as games with strategically rational players. Similar points apply to the way political scientists think about parties and other organizations in politics.

The following argument summarizes the case for realism about group agency:[5]

Premise 1: The ascription of intentional agency to certain organized collectives is explanatorily indispensable if we wish to make sense of their behaviour.

Premise 2: If the ascription of some property to an entity is indispensable for explaining that entity's behaviour, then we have a provisional justification for assuming that the entity really has that property.

[4] Recall the references in footnote 1.
[5] For an earlier version of this argument, see List (2018). The argument is also implicit in other works on group agency, including Tollefsen (2002) and List and Pettit (2011).

Conclusion: We have a provisional justification for assuming that the collectives in question really have the property of intentional agency.

Premise 1 is a partly empirical and partly methodological claim about the social sciences. Premise 2 follows from a broader philosophical principle known as the "naturalistic ontological attitude".[6] That principle asserts that when we seek to answer ontological questions about which entities and properties are real in any given domain, we should consult our best scientific theories of that domain and be guided by considerations of explanatory indispensability. For example, the fact that postulating gravity and electromagnetic fields is indispensable for explaining certain physical phenomena gives us a provisional justification for thinking that gravity and electromagnetism are real. The justification is provisional, because there might still be special reasons for doubting our best scientific theories in the relevant domain, and those theories remain open to revision. But, in the absence of any contrary reasons, we are warranted in taking the ontological commitments of our best theories at face value. If we accept Premises 1 and 2, then we must recognize group agency as a real phenomenon, at least provisionally.

To be sure, the present argument supports realism about group agency only for those collectives for which the ascription of agency is explanatorily indispensable. Firms and states are plausible examples. Unorganized collectives and random collections, such as the collection of pedestrians on the street at this moment, would not qualify, as we would not get any explanatory benefit from ascribing agency to them. Similarly, not every case of cooperative or coordinated activity produces a group *agent*. When workers go on strike or villagers cooperate in providing some public good, the *locus* of agency remains the individual, even if the participants jointly bring about some aggregate effect, receive some gains from cooperation, and/or share certain intentions. Such collective or joint actions may be *ingredients* in the implementation of group agency (since organized collectives often rely on collective or joint action), but collective or joint action alone is insufficient for group *agency*.[7]

3. Artificial intelligence

Artificial intelligence (*AI*) is the increasingly ubiquitous phenomenon that some artificial (e.g., computational or robotic) systems display certain cognitive capacities and/or autonomous agency that are in some respects analogous to the cognitive capacities and/or agency of humans

[6] See, e.g., Quine (1977) and Fine (1984).
[7] On collective action, see, e.g., Olson (1965) and Ostrom (1990). On joint action, see, e.g., Bratman (1999, 2014), Gilbert (1989), and Tuomela (2007). On norm-governed multi-agent interactions, see also Boella and van der Torre (2007).

or other animals. Defining artificial intelligence by reference to cognition or agency is very common. For instance, Stuart Russell and Peter Norvig define the subject of AI as "the study of agents that receive percepts from the environment and perform actions".[8] Depending on which cognitive and agential capacities we focus on, we get more or less demanding definitions.

Examples of AI systems range from driverless cars, autonomous air vehicles (drones), medical helper robots, diagnostic devices, and financial trading systems, all the way to chess-playing computers, navigation devices, sophisticated vending machines, and even robots that can autonomously assemble IKEA furniture.[9] The first few examples display greater autonomy and flexibility than the last few, which are more specialized.

A familiar distinction is that between "weak" and "strong" AI. "Strong AI" refers to machine intelligence comparable in sophistication, flexibility, and/or generality to human intelligence. "Weak AI" refers to machine intelligence "weaker" than human intelligence, for instance in the sense of being more specialized and less flexible. Strong AI on a par with human intelligence is not yet a reality (and a controversial notion), and it is not clear how soon it is likely to be developed, if at all.[10] The term "artificial superintelligence", finally, refers to (so far, extremely hypothetical) AI systems whose capacities exceed human ones.[11]

Even though human-like AI is still elusive, the cognitive and agential capacities of AI systems are getting more sophisticated. As such systems attain greater autonomy and decision-making power, they raise new ethical and regulatory challenges. While early AI systems had only limited capacities and tended to be employed only in very controlled environments, we increasingly face the arrival of AI systems that make high-stakes decisions and operate more autonomously in less controlled environments.[12] If a system has only limited capacities, such as a robotic floor cleaner or a pre-programmed factory robot, or if its use has no serious spill-over effects beyond a restricted environment, as in the case of an automated train in a tunnel, then it does not give rise to *qualitatively novel* risks, compared to earlier technologies. Of course, the large-scale use of such systems may raise legitimate concerns about automation-driven job losses, but labour-market implications of new technologies are not unprecedented. By contrast, if an AI system operates relatively freely in a largely uncontrolled environment,

[8] See Russell and Norvig (2009). An account of AI in terms of agency can also be found in Laukyte (2014, 2017).
[9] On the last example, see, e.g., Chokshi (2018).
[10] Arguably, intelligence is best understood, not as one-dimensional, but as a cluster of capacities, with multiple, perhaps independent components: mathematical-logical, linguistic-verbal, musical, artistic-creative, motor-control, mind-reading etc. In line with this, Pinker (2018, p. 298) describes the concept of artificial *general* intelligence as "barely coherent", also quoted by Aaronson at <https://www.scottaaronson.com/blog/?p=3654>.
[11] On artificial superintelligence, see, e.g., Chalmers (2010) and Bostrom (2014).
[12] For discussion, see, e.g., Fisher, List, Slavkovik, and Winfield (2016).

as in the case of a driverless car or a fully autonomous drone, or if it can make high-stakes decisions on its own, as in the case of some medical, financial, and military systems, then the societal implications are qualitatively novel. We are then dealing with *artefacts as genuine decision-makers*, perhaps for the first time in human history. My focus in this paper is on such systems, which can make decisions with non-trivial stakes and operate relatively autonomously. The challenges raised by systems with more limited capacities and less autonomy are not my topic here, as they are less novel. Moreover, some of those weaker systems might be best viewed as advanced tools, not as genuinely autonomous agents.

4. The key parallel

The key parallel between group agency and artificial intelligence, as anticipated, lies in the fact that both phenomena involve entities distinct from individual human beings that qualify as goal-directed, "intentional" agents, with the ability to make a significant difference to the social world.[13] As we will see later, those entities are not thereby guaranteed to qualify as *moral* agents too, which is a more demanding requirement than intentional agency alone.

To develop the parallel, let me give a more precise definition of intentional agency. An *intentional agent* is an entity, within some environment, that meets at least three conditions:[14]

- It has representational states, which encode its "beliefs" about how things are.
- It has motivational states, which encode its "desires" or "goals" as to how it would like things to be.
- It has a capacity to interact with its environment on the basis of these states, so as to "act" in pursuit of its desires or goals in line with its beliefs.

To give a simple example: I believe that there is tea available in the kitchen; I desire to drink tea; I then act by going to the kitchen to get some tea. My action is explained – in fact, rationalized – by my beliefs and desires, and this pattern applies more generally.

Consistently with these conditions, agents can vary considerably in their capacities and sophistication. Their cognitive make-up can range from simple to complex. Human beings, the most familiar agents, have many psychological states beyond beliefs and desires. These include emotions, hopes, fears, and various subconscious mental states. Adult human beings also have the capacity for normative reasoning, on top of their more mundane capacities. But although human beings stand out, entities ranging from non-human animals to autonomous robots and

[13] As noted, this parallel is also discussed in Laukyte (2014, 2017).
[14] I here build on List and Pettit (2011, ch. 1) and List (2018). The notion of a "belief-desire agent" goes back to David Hume. For an influential account of "belief-desire-intention agency", see Bratman (1987).

organized collectives can meet the agency conditions too. Behavioural ecologists often study animal behaviour using game theory and other theories involving the attribution of goal-directed agency to the animals in question. And, as already noted, robots and AI systems, as well as corporate entities, are routinely viewed as agents in the relevant fields of study. So, we can certainly think of AI systems and group agents as intentional agents of a non-biological sort.[15]

The parallel is further reinforced if we look at the most famous test for artificial intelligence, the "Turing Test", and compare it with a prominent test for intentional agency, the "intentional stance" test, proposed by Daniel Dennett, which is often also applied to organized collectives. Let's begin with the Turing Test. Alan Turing argued that, to determine whether a system is "intelligent", we should ask whether its behaviour leads a human observer to believe that it has human-like cognitive and agential capacities.[16] Specifically, Turing suggested that we should replace the hard question "Can machines think?" with the more tractable question "Are there imaginable digital computers which would do well in the imitation game?", where this means displaying behaviour that is indistinguishable, in an observer's eyes, from that of an intelligent agent.[17] Different specifications of the imitation game correspond to different versions of this test. In Turing's own version, the system must imitate the responses of a human interlocutor in written communication via an instant-messaging system. But other, less demanding versions are also conceivable. A Turing Test for self-driving cars might focus on whether a self-driving car can imitate the behaviour of a human-operated car, even in complex traffic situations.

Let's compare this with the "intentional stance" test for agency.[18] For Dennett, the criterion for whether an entity is an intentional agent is whether we can make sense of its behaviour by *viewing* it as an intentional agent. As he puts it, "[a]nything that is usefully and voluminously predictable from the intentional stance is, by definition, an intentional system."[19] Taking the "intentional stance" towards an entity means interpreting it *as if* it were an intentional agent and then explaining and predicting the entity's behaviour on that basis. If this strategy works, the entity qualifies as an intentional agent; if not, then not. In the case of most inanimate systems, such as the solar system, a volcano, or the earth's atmosphere, Dennett's criterion is not met. We are better off taking a "physical stance" towards them: explaining them as systems governed by physical laws of nature and causal mechanisms. In the case of a human being or complex animal, however, Dennett's criterion is met and vindicates the presence of intentional

[15] For a useful discussion, see also Dretske (1999).
[16] See Turing (1950).
[17] Ibid., p. 442.
[18] See Dennett (1987).
[19] See Dennett (2009, p. 339).

agency. Similarly, as theorists of group agency have argued, firms, corporations, and other organized collectives are "usefully and voluminously predictable from the intentional stance".[20]

To see the similarity between the "intentional stance" test for agency and the Turing Test for intelligence, consider the relationship between

(i) successful interpretability as an agent, and

(ii) doing well in a corresponding suitably specified imitation game.

One may expect (i) to be correlated with (ii). Any entity that is successfully interpretable as an agent is likely to do well in an imitation game designed to test for the relevant agential capacities, and any entity that does well in this imitation game is likely to be successfully interpretable as an agent.

While both Turing and Dennett, in effect, propose *interpretivist* accounts of agency, I prefer a stronger *realist* account. On an *interpretivist* account, to be an agent *is* to be successfully interpretable as an agent. For me, interpretability as an agent is *evidence* for agency, but not *conceptually the same* as agency. On my stronger *realist* understanding, to be an agent is to satisfy the three conditions for agency introduced above. So, while from an interpretivist perspective successful interpretability as an agent is *constitutive* of agency, for me it is merely *indicative*. That an entity is interpretable as an agent is good *evidence* for the hypothesis that it satisfies the agency conditions. Irrespective of whether we go with an interpretivist criterion or my preferred realist one, however, the fact that sophisticated AI systems and suitably organized collectives meet criteria (i) or (ii) supports the claim that they each qualify as agents.[21]

In fact, group agents can be viewed as special cases of AI systems, where the "hardware" supporting their artificial intelligence is social rather than electronic. Rohit Parikh has coined the term "social software" to capture the idea that certain social phenomena can be understood as computational processes implemented in social systems. Parikh uses this term to refer to a broad range of social phenomena, such as procedures and institutions, voting systems, and other mechanisms, and he does not focus on group agency.[22] But, consistently with Parikh's usage, the idea that group agents are socially implemented AI systems seems quite natural. As functionalist philosophers of mind have long argued, agency is a multiply realizable phenomenon: it can be realized by different "hardware systems". The foregoing discussion illustrates that agency admits at least three different kinds of hardware: biological, as in the case of human beings and non-human animals; electronic, as in the case robots and other AI

[20] See Tollefsen (2002, 2015) and List and Pettit (2011).
[21] Tollefsen (2015) defends group agency from an interpretivist perspective.
[22] See Parikh (2002).

systems; and social, as in the case of group agents. The boundaries between these three kinds of hardware may become increasingly blurred with the arrival of new (often controversial) technologies, from biological robotics and human enhancement to technologically augmented social systems.

While the main parallel between group agency and artificial intelligence consists in the instantiation of agency beyond the human individual, there are other parallels too. Group agents and AI systems each have an internal architecture – social in the one case and computational in the other[23] – and it may be illuminating to compare these architectures, perhaps applying the computational paradigm suggested by Parikh's notion of social software. For instance, group agents and AI systems each tend to receive and aggregate inputs from different sources. To function well, they each need to satisfy certain rationality conditions at the level of the entity as a whole. They must each revise or update their beliefs in response to input from the environment. And so on. But these more technical parallels are not my focus here, and I just flag them as issues for further investigation. I want to move on to the moral and regulatory challenges raised by group agents and AI systems.

5. Challenges related to responsibility

5.1. Responsibility gaps in group agency and AI

Group agents have considerable power and influence in the world and can cause significant harms. States can wage wars and do many things that affect the lives of their citizens and others. Likewise, corporations can affect people's well-being through their business decisions and sometimes even cause disasters. Think of the Deepwater Horizon oil spill in the Gulf of Mexico in 2010, caused by the drilling operations of a multinational corporation, or the deforestation caused by big industry. As is widely recognized, in the case of such harms, it is sometimes difficult to identify one or several individuals to whom all of the relevant responsibility can reasonably be attributed. The sum-total of individual responsibility may intuitively fail to do justice to the full amount of the harm caused. There is then a shortfall of attributed responsibility: a "responsibility gap" or "responsibility void".[24]

[23] In the case of group agency, French (1984) speaks of the "corporate internal decision structure" and List and Pettit (2011) speak of the group's "organizational structure", with a special emphasis on aggregation functions. Parallels between the constitution of a social entity and that of a mind were also discussed by Minsky (1986).

[24] See, among others, Braham and van Hees (2011), Collins (2017, 2018), and Duijf (2018). On corporate responsibility more generally, see French (1984), Erskine (2001), Copp (2006), Pettit (2007), and List and Pettit (2011, ch. 7), among others. For an overview, see Smiley (2017).

Philip Pettit gives the example of the Herald of Free Enterprise, a passenger ferry which sank in the English Channel in 1987, killing almost two hundred people.[25] An inquiry concluded that the ferry company was very sloppy and had terrible safety procedures. In a commentator's words, "[f]rom top to bottom the body corporate was infected with the disease of sloppiness".[26] Yet, despite a complex legal aftermath, no-one was found legally responsible to an intuitively appropriate extent. There was a shortfall in legally attributable responsibility, since the ferry company was not treated as a *locus* of responsible agency.[27] The focus on individual responsibility meant that the systemic failure at the collective level fell below the law's radar screen.

What should we say in such cases? One possibility is to deny that there is any form of responsibility other than individual responsibility. We would then have to do our best to identify all the individuals who can be held responsible for at least parts of the harm caused, but if the sum-total of their responsibility falls below the amount of responsibility we would intuitively like to attribute, we would have to accept that there is no more responsibility to be assigned. The situation would be akin to a natural disaster or fluke accident, where despite the regrettable harm no responsibility can be attributed to anyone, aside from possible responsibility for damage prevention and disaster response.

However, there is an important difference between natural disasters and corporate harms. While in a natural disaster the relevant harms cannot normally be traced back to intentional actions, corporate harms have an agential source: they stem from the actions of a group agent. For this reason, theorists of group agency often argue that we should treat corporate entities as *loci* of responsibility themselves, over and above their members. In the example of the ferry company, the responsibility gap might have been avoided by holding the company itself responsible for its sloppy procedures. The assignment of corporate responsibility may be combined with appropriate sanctions at the corporate level: targeting, for instance, the assets of the corporation or its operating permissions.

Let me now turn to the case of AI. As AI systems are becoming increasingly ubiquitous as well as powerful, there are bound to be some responsibility gaps too. If AI systems operate autonomously in high-stakes settings, in areas ranging from transportation and medicine to finance and the military, the occurrence of some harms is inevitable in practice.[28] Just think of

[25] See Pettit (2007) and also List and Pettit (2011, ch. 7).
[26] See Colvin (1995), quoted also in Pettit (2007).
[27] Ibid.
[28] See, e.g., Matthias (2004), Sparrow (2007), and Broersen (2014).

all the things that might go wrong with self-driving cars, military drones, autonomous financial trading systems, and diagnostic systems or helper robots in medicine. In the case of such harms, we cannot assume that full responsibility can always be attributed to human operators, developers, owners, or regulators. Rather, it is conceivable that, in some cases, the sum-total of human responsibility will intuitively fail to do justice to the full amount of the harm caused. If this happens, we are faced with a responsibility gap. Perhaps all humans involved – the system's operators, developers, owners, and regulators – were sufficiently diligent and acted conscientiously, and nonetheless the AI system, perhaps due to some unforeseeable dynamics, caused a harm. A high level of complexity and autonomy will almost inevitably go along with a certain degree of unpredictability. Andreas Matthias describes the problem as follows:

> "Traditionally, the manufacturer/operator of a machine is held (morally and legally) responsible for the consequences of its operation. Autonomous, learning machines, based on neural networks, genetic algorithms and agent architectures, create a new situation, where the manufacturer/operator of the machine is in principle not capable of predicting the future machine behaviour any more, and thus cannot be held morally responsible or liable for it. The society must decide between not using this kind of machine any more (which is not a realistic option), or facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription."[29]

Although it is contentious whether unpredictability alone is enough to absolve manufacturers and operators of responsibility (I would argue it is not), AI-related responsibility gaps nonetheless seem possible, and so we need to take a stand on them. Should we insist that, beyond the responsibility that can reasonably be assigned to the human operators, owners, manufacturers, and regulators, there is no further responsibility to be assigned? On this view, the resulting harms would be on a par with those caused by a natural disaster or fluke accident. Or should we attribute responsibility to the AI systems themselves? At first sight, the idea of assigning responsibility to an AI system may seem far-fetched and unduly anthropomorphist, but I want to suggest – in line with the views of Laukyte, Solum, and others – that AI responsibility, like corporate responsibility, might be defensible.[30]

[29] See Matthias (2004, p. 175).
[30] See, in particular, Laukyte (2014, 2017) and Solum (1992), as cited earlier.

5.2. Why corporate and AI responsibility cannot be reduced to human responsibility

Before explaining the case for AI responsibility on the model of the more familiar notion of corporate responsibility, it is worth addressing the objection that each of these purported forms of responsibility can be reduced to human responsibility. Let's begin with corporate responsibility. Group agents are collections of individuals. A group agent cannot act unless some of its members act; what a group agent does is determined by what the members do, under the relevant organizational structure.[31] Since the members are themselves responsible agents, so the objection goes, they must therefore be responsible for the group's actions.

This objection, however, misses a key conceptual point about group agency. In the case of genuine group agency, as opposed to mere collective or joint action, the collective constitutes a new *locus* of agency, distinct from the agency of any of the individual members. The fact that those individuals *causally* or even *constitutively* contribute to the group's corporate actions does not imply that those actions can also be *agentially* attributed to them. If we take the logic of group agency seriously, we must not treat the actions of the group agent as the actions of its members. It is a contingent matter whether, and how, the members are involved in a group agent's actions *qua responsible agents* and not merely *qua causal contributors* or *qua ingredients in the group's organizational structure*, perhaps even unwittingly or involuntarily. Arguably, the members can be held responsible for a group agent's actions only to the extent that they have played certain normatively relevant roles:[32]

- an *enacting role*, for instance as a knowing, willing, and uncoerced manager, official, or representative of the group agent,
- an *authorizing role*, for instance as a director, board member, owner, share-holder, or regulator, or
- an *organizational-design role*, for instance as a founder, framer, policy maker, or institutional designer.

There is no reason to expect that the sum-total of the members' responsibility due to those roles will always be commensurate with the total amount of corporate responsibility we might ideally like to attribute. Thus, responsibility gaps remain possible.

Let me now turn to AI responsibility. Here one might raise a similar objection. One might argue that, since AI systems are always developed, built, owned, and operated by humans, AI responsibility must ultimately be reducible to human responsibility. Again, the objection

[31] On this "determination" or "supervenience" thesis, see List and Pettit (2006, 2011).
[32] See also List and Pettit (2011, ch. 7) and List (2019).

misses a key point, namely that, no matter how AI systems have been brought into existence, systems above a certain threshold of autonomy constitute new *loci* of agency, distinct from the agency of any human designers, owners, and operators. In fact, such systems can arguably become even more autonomous than group agents. Setting aside hybrid systems such as drones operated by remote human pilots, AI systems do not require human participation in the same way in which group agents do. Once in operation, AI systems can potentially operate with little or no human input. Furthermore, approaches such as evolutionary computing might bring into existence AI systems that have no human designer or operator at all: they may have evolved in an artificial environment. And so, there is no conceptual reason to think that the responsibility of AI systems will always be reducible to human responsibility.

Think of an analogy: a person's parents play a key causal role in making him or her the person he or she is, but the adult human being is nonetheless an agent distinct from his or her parents, and parents cannot normally be held responsible for their adult children's conduct. Likewise, from the fact that someone has played a role within the causal history of an AI system, it does not automatically follow that this person is responsible for the system's conduct. Rather, it is a contingent matter to what extent other agents – individual or corporate – are responsible for what an AI system does. The answer will vary from case to case. As in the case of group agency, human responsibility may plausibly be assigned for an AI system's actions only to the extent that the relevant people have played one of the following roles:

- an *enacting role*, for instance as a knowing, willing, and uncoerced operator of the system,
- an *authorizing role*, for instance as an owner, provider, or regulator, or
- a *system-design role*, for instance as a creator, manufacturer, software designer, or policy maker.

Crucially, there is no guarantee that the sum-total of any such agents' responsibility will always be commensurate with the total amount of responsibility we might ideally like to attribute for an AI system's actions. Therefore, we cannot rule out the possibility of responsibility gaps, as Andreas Matthias has pointed out.[33]

5.3. Fitness to be held responsible in group agents and AI systems

Attributing responsibility to corporate or artificial entities themselves would enable us to avoid certain responsibility gaps. But this fact alone is insufficient to establish that there truly is

[33] Recall Matthias (2004).

corporate or AI responsibility. It would be a mistake to derive the conclusion that an entity can be assigned responsibility simply from the premise that such an assignment helps us to avoid responsibility gaps. Even if a hurricane causes a huge amount of damage for which there is little human responsibility, it does not follow that the hurricane itself is a bearer of responsibility. Only entities with the right capacities can be bearers of responsibility. So, before we can assign responsibility to an entity, we must establish that this entity is of the right sort.

What conditions must an entity meet to be "fit to be held responsible"?[34] Although there is some room for debate here, the following seem plausible conditions:[35]

> **Moral agency:** The entity is an agent with the capacity to make normative judgments about its choices – judgments about what is right and wrong, permissible and impermissible – and to respond appropriately to those judgments.
>
> **Knowledge:** The entity has the information needed for assessing its choices normatively, or at least reasonable access to such information.
>
> **Control:** The entity has the control required for choosing between its options.

For the sake of my argument, I will assume that these conditions, perhaps after suitable fine-tuning, are necessary and sufficient for fitness to be held responsible. With regard to the first condition, note that intentional agency alone, without the capacity for normative judgments, is not enough for fitness to be held responsible. Non-human animals, for instance, are intentional agents but lack moral agency, and indeed they cannot be bearers of responsibility. The second condition rules out cases in which agents, through no fault of their own, lack certain information that would be needed for a normative assessment of their choices. Agents who are brainwashed or trapped in an informationally deprived environment may be examples. They wouldn't bear responsibility – certainly not full responsibility – for choices made under such constraints. The third condition, finally, rules out responsibility in cases where agents lack control over their choices, due to external or psychological factors. If, however, an entity meets all three conditions, as in the case of a typical adult human being, there seems no reason to refrain from treating that entity as a bearer of responsibility.

In general, it is an empirical question whether a given entity – biological, social, or electronic – meets the three conditions. Although, in the biological case, humans are the only

[34] I am borrowing this term from Pettit (2007).
[35] These conditions (also discussed in List 2019) are adapted from List and Pettit (2011, ch. 7), drawing, in turn, on Pettit (2007). Laukyte (2014, 2017) also employs variants of these conditions to argue for AI responsibility.

known entities satisfying them, there is no conceptual reason why non-human entities could not satisfy them too. In the case of group agents, it is widely accepted (especially since Peter French's influential work on corporate responsibility) that at least some corporate entities are fit to be held responsible.[36] Whether a particular organized collective, say a commercial corporation, meets the relevant requirements depends on how it is set up and organized.

One crucial issue is whether the entity is not just an *intentional* agent but meets the stronger requirements for *moral* agency. Well-organized corporations have procedures and mechanisms in place that allow them to make corporate-level judgments about what is permissible and impermissible, and to act on those judgments. Indeed, many organizations have compliance departments and ethics committees.[37] Such entities are what Philip Pettit calls "conversable":[38] we can engage with them and challenge their actions on the basis of normative reasons, not unlike the way we engage with adult human beings. This is very different from the way we engage with, say, a non-intentional physical process or a non-human animal. The latter kinds of entities are not capable of normative reasoning. We can *causally* interact with them, but we cannot influence them *by giving them normative reasons*, and they would not be able to give us reason-based justifications for their behaviour.

The present observations suggest a plausible regulatory requirement for the creation and operation of powerful group agents in society.

> **A proposal:** Society, via its regulatory authorities, should permit the creation and operation of powerful group agents, such as corporations and other organizations in high-stakes settings, *only if* structures are in place to ensure their fitness to be held responsible for their corporate actions.[39]

The idea is that an organization, if operative in a "high-stakes setting" according to society's criteria, would attain its official incorporated status and operating licence only if it can show that it has procedures and mechanisms in place that ensure that it satisfies the three conditions stated above and that it therefore qualifies as a responsible moral agent. Depending on what is at stake, society may further impose precise restrictions on what the organization is or is not allowed to do and demand that the organization should have appropriate financial assets and/or

---

[36] Recall the references in note 6.
[37] For recent discussions of corporate moral agency, see Björnsson and Hess (2017) and Pasternak (2017).
[38] See, e.g., Pettit (2001).
[39] Along similar lines, a "developmental rationale" for holding group agents responsible is defended in List and Pettit (2011, ch. 7).

insurance arrangements such that it can be fined or made to pay compensation in cases of wrongdoing. In this way, society can protect itself against problematic responsibility gaps.

Moving on to artificial intelligence, I suggest that, here too, it is an empirical question whether a given system satisfies the conditions for fitness to be held responsible. Obviously, first-generation AI systems do not come close, and even if goal-directed behaviour is common in AI systems, there is still a big gap between mere intentional agency and moral agency. The emerging research programme of "engineering moral machines" is an attempt to implement moral agency in AI systems: to develop AI systems that behave morally.[40]

There is a pressing need to ensure that AI systems do not just pursue certain pre-programmed goals mechanically, but that they comply with moral norms. It is inevitable that autonomous systems in high-stakes settings will sometimes be confronted with decisions requiring moral judgment. For a simple example, think of a self-driving car that has a choice between hitting a child on the road or swerving with the consequence of damaging the mirror of a parked vehicle or mildly injuring its own passengers. Given the vast number of different possible such situations, it is not feasible to provide the car with a complete list of definitive instructions in advance. Moral guidelines cannot be exhaustively enumerated like this. Rather, the car must have the capacity to assess the relevant situations autonomously and to arrive at reasonable moral decisions on its own. In short, it must engage in normative reasoning and achieve a certain form of moral agency.

It should be clear that, while there are significant technical challenges here, conceptually, there is no reason why an AI system could not qualify as a moral agent and, in addition, satisfy the knowledge and control conditions I have stated. Even if existing AI systems do not yet meet these requirements, there is no reason to think that having an electronic or otherwise engineered hardware is an in-principle barrier to their satisfaction. If we acknowledge that group agents can qualify as fit to be held responsible, then we should be prepared to acknowledge that AI systems can do so too, at least in principle. In fact, if I am right that group agents can be viewed as socially implemented AI systems, then the familiar notion of corporate responsibility can already be viewed as an existing example of AI responsibility.

In analogy to the regulatory requirement that I have endorsed in the case of group agents, one might propose a similar requirement for AI systems, albeit in slightly amended form:

[40] See, e.g., Broersen (2014) and Fisher, List, Slavkovik, and Winfield (2016). For an overview of the debate on artificial moral agency, see Fossa (2018).

> **A proposal:** Society, via its regulatory authorities, should permit the use of autonomous AI systems in high-stakes settings *only if* structures are in place to ensure these systems' – or at least their legal representatives' – fitness to be held responsible for their actions.

The idea is that the operating license for any AI system in what is deemed a "high-stakes setting" would be contingent on evidence that the system meets the conditions for fitness to be held responsible and/or that there is a full transfer of responsibility to certain legal representatives. The proposed form of AI responsibility may, in turn, have to be underwritten by certain assets, financial guarantees, and/or insurance, so that, in the event of a harm, the system or its legal representatives can be made to pay appropriate fines and compensation. Of course, society would have to decide what counts as a "high-stakes setting" for regulatory purposes. The use of powerful AI systems in such settings would then be prohibited by the relevant regulators unless these systems properly function as *loci* of responsible agency themselves or there is a full transfer of responsibility to their human legal representatives as a backup.[41]

As in the case of group agents, this regulatory approach would enable society to protect itself against responsibility gaps. It is important that we avoid a situation in which individuals or corporate entities can evade liability for high-risk decisions by delegating those decisions to AI systems and then hiding behind the autonomy and unforeseeability of those systems' behaviour. In a slogan, it should not be possible to achieve impunity for harmful actions just because "the algorithm did it".

To implement moral agency in AI systems, we would have to design those systems explicitly to have ethical decision-making capacities, in the same way in which corporations may be required to have ethical compliance mechanisms. This can be technically achieved in at least two ways: *either* by pre-programming certain moral constraints into the AI system *or* by training the system to recognize morally significant situations and to adjudicate them autonomously.[42] The first, constraint-based approach would require the codification of the relevant moral norms in a machine-implementable format and programming those norms explicitly into the AI system. The second, training-based approach would require the use of

---

[41] In line with this, Sparrow (2007) argues against the use of certain forms of military AI technology. He writes: "the more autonomous these systems become, the less it will be possible to properly hold those who designed them or ordered their use responsible for their actions. Yet the impossibility of punishing the machine means that we cannot hold the machine responsible. We can insist that the officer who orders their use be held responsible for their actions, but only at the cost of allowing that they should sometimes be held entirely responsible for actions over which they had no control. For the foreseeable future then, the deployment of weapon systems controlled by artificial intelligences in warfare is […] unfair either to potential casualties […] or to the officer who will be held responsible for their use" (pp. 74–75).

[42] For the distinction, see Fisher, List, Slavkovik, and Winfield (2016, p. 468).

machine-learning techniques to recognize patterns in a database of illustrative moral decisions and to extrapolate those patterns to future decisions, just as AI systems can learn to recognize patterns on chest x-rays, given a sufficiently large database of images paired with corresponding diagnoses (an instance of "supervised learning"). For instance, a self-driving car might be trained using a database of thousands of traffic situations paired with information about the morally desirable response in each situation.

Both approaches have advantages and disadvantages. Under the constraint-based approach, it may be easier to verify in advance that the system will be norm-compliant, but as moral philosophers recognize, the systematic codification of moral norms is extremely difficult. Under the training approach, it may be easier to emulate human moral judgments, provided the training database is large enough, but it is harder to verify the system's reliability in advance. In short, there may be a trade-off between *verifiability* and *moral adequacy*. More research is needed to make progress here.

A further strategy, beyond engineering AI systems to behave morally, is to require that AI systems be equipped with an "ethical black box", as proposed by Alan Winfield and Marina Jirotka.[43] This is "the equivalent of a Flight Data Recorder to continuously record sensor and relevant internal status data". It enables investigators to reconstruct the AI system's internal decision-making process in the context of any accident.[44] Winfield and Jirotka argue that this is "an essential part of establishing accountability and responsibility" in robots and autonomous systems, and that "without the transparency afforded by an ethical black box, [such] systems are unlikely to win public trust".[45]

Finally, in those cases in which it is unrealistic to turn AI systems into genuinely responsible agents but we still wish to use such systems, it may be appropriate to introduce a regime of strict liability for their operation, akin to strict-liability regimes in other industries, such as the food industry. Under a strict-liability regime, the owners or operators of any AI system would be liable for any harms done by the system, even if no fault, negligence, or intention by those owners or operators can be established. A strict-liability regime might plausibly incentivize the development of AI systems that are as safe as possible, and it would be likely to prompt owners and operators to purchase insurance to cover any residual risks. However, it is important for society to decide if (and when) AI systems should ever be admitted as genuine decision-makers in high-stakes settings if they do not display at least a modicum of moral agency.

---

[43] See Winfield and Jirotka (2007).
[44] Ibid.
[45] Ibid.

The bottom line is that, in the case of AI as much as in the case of group agency, careful regulation is needed to avoid situations in which responsibility gaps can occur by accident or in which risky decisions are delegated to non-human decision-makers in order to minimize human responsibility for any harms. AI systems that qualify as responsible moral agents, like corporate moral agents, might, in turn, achieve what Philip Pettit calls "conversability": we would be able to scrutinize and seek justifications for their actions through the lens of normative reasons, rather than merely engage with them in a causal or mechanical manner.

6. Challenges related to rights and moral status

6.1. Corporate and AI rights and legal personhood

If we regard some corporate entities and AI systems as moral agents who are fit to be held responsible, should we also give them rights?[46] Should we agree, for instance, with the Citizens United decision of the US Supreme Court, which gives relatively broad free-speech rights – exercisable via monetary spending – to commercial corporations and other organizations? And should we then extend those rights to AI systems, a serious question in a world of Twitter bots and automated contributors to social media feeds? Moreover, should we treat group agents and AI systems as persons of their own, akin in some respects to human persons?

At first sight, the idea of giving rights not just to corporations but also to AI systems may seem preposterous, and the idea that such entities could attain some kind of moral status seems even more far-fetched. However, consider the functions performed by corporate entities in modern societies: the state provides vital services, law-enforcement agencies uphold the law, universities educate people, and corporations and banks – despite some of their negative aspects – play important roles in the economy. It is widely recognized that those entities would not be able to perform those functions if they did not have at least some rights (carefully circumscribed rights, of course) and some associated legal status: the right to own property, the right to employ people and to enter contracts, the right to demand the fulfilment of certain obligations by others and to sue non-compliant parties in court, and so on. Indeed, corporate entities are routinely considered *legal* persons, and it is this status that qualifies them as bearers

[46] For earlier discussions of these questions in the corporate case, see, e.g., List and Pettit (2011), and in the AI case, see Solum (1992), Laukyte (2014, 2017), and Turner (2018). Solum and Turner discuss the case for legal personhood for AI, noting the parallel with corporations, and Laukyte argues that the "performative" (or functionalist) case for granting personhood to group agents considered in List and Pettit (2011) essentially carries over to AI systems.

of legal responsibilities and as holders of certain legal rights. To be sure, legal personhood is not the same as moral personhood, but it is an important status nonetheless.

Analogously, we may wonder whether a similar legal status should be extended to AI systems, as they play increasingly important roles in our complex world. Perhaps some of those systems, too, will be able to perform some of their functions – especially future functions – only if they have a certain legal status and even certain rights, a possibility already entertained by Solum in the early 1990s.[47] Presumably, we need to regulate the decision-making powers of AI systems in financial, medical, and military contexts. We need to specify, for instance, whether autonomous systems in business contexts may enter valid contracts, whether medical systems may prescribe medicines and perform treatments, and whether military systems may make certain decisions autonomously in critical situations.

Along these lines, a 2017 European Parliament report on legal and ethical aspects of robotics recommends that the European Commission should explore the possibility of

> "creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently".[48]

Supporters of this proposal suggest that it would be appropriate in light of the ever-growing range of functions performed by autonomous systems in modern societies. As a commentator puts it, the proposed status "could allow robots to be insured individually and be held liable for damages if they go rogue and start hurting people or damaging property." And further: "Legal personhood would not make robots virtual people who can get married and benefit from human rights …; it would merely put them on par with corporations, which already have status as 'legal persons,' and are treated as such by courts around the world."[49] Critics object that the proposal is premature, and that "[b]y seeking legal personhood for robots, manufacturers were merely trying to absolve themselves of responsibility for the actions of their machines".[50]

However, even if we gave legal personhood to AI systems, this would not absolve relevant humans – operators, manufacturers, owners, or regulators – of responsibility any more than

---

[47] See Solum (1992).
[48] See <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN>.
[49] See Delcker (2018). On AI legal personhood, see also Turner (2018).
[50] This point was made by Noel Sharkey, quoted in Delcker (2018).

giving legal personhood to corporations absolves directors, managers, and relevant employees of individual responsibility. Even if certain corporate or electronic entities are treated as legal persons, human beings may still be held responsible to an appropriate extent for those entities' actions – namely to the extent that they have played certain normatively relevant roles, such as enacting, authorizing, or design roles, as discussed earlier.

To make progress in the debate on whether corporate rights and personhood should be extended to AI systems, we need to say more about what granting rights and/or personhood to an entity means and what the justification could be.

6.2. Rights and personhood, legal and moral

Whenever we speak of "rights", we must clarify at least two things. First, we must clarify whether we are referring to legal rights or to moral ones. To illustrate this distinction, note that people living under a dictatorial regime have far weaker legal rights than people living in a liberal democracy, but we think that their moral rights are the same, even if these moral rights are not respected in their society. Second, we must clarify whether we are referring to rights in a thin sense, understood simply as (i) *deontic powers or permissions* to do or claim certain things, or to rights in a richer sense, understood as (ii) a *status* in which those deontic powers or permissions are grounded and by reference to which they are justified.[51] Let me call (i) the "deontic" sense of rights, and (ii) the "status" sense. Saying that someone has a right in the deontic sense is just another way of saying that this person has certain deontic powers or permissions. It carries no implication as to what the *grounds* of those powers or permissions are and how they are *justified*. By contrast, when we are referring to a right in the status sense, for instance, a *natural* right, a *human* right, or an *animal* right, then we are saying something about the grounds, or the justification, of the relevant powers or permissions.

Similarly, the term "personhood" can be used in different senses. It can stand either for a legal status or for a moral one. Legal personhood consists in having a certain package of legal powers, permissions, and responsibilities, but it need not come with any special moral status. Legal persons need not have any intrinsic moral significance. Moral personhood, by contrast, is not a legal status but a moral one. It entails a package of moral powers, permissions, and responsibilities, and comes with a certain intrinsic moral significance. Moral persons are objects of moral concern; they matter intrinsically.

[51] For a closely related distinction, see Valentini (2018).

With these clarifications in place, we can see that, even if we grant legal personhood and legal rights to corporate entities and/or AI systems, this does not imply granting them moral personhood and moral rights. Moreover, giving certain deontic powers and permissions to corporate or artificial entities does not imply that these entities have intrinsic moral significance. They might be granted those powers and permissions for purely instrumental reasons, for instance because this helps them to perform certain useful functions in society. And they might be denied other powers and permissions, also for instrumental reasons.

Let me use the term "non-derivative rights" to refer to deontic powers or permissions (whether legal or moral) that are justified by reference to an entity's intrinsic moral significance, as in the case of human rights and (arguably) animal rights, and let me use the term "derivative rights" to refer to deontic powers or permissions that are justified in more indirect, typically instrumental way, for instance by reference to some other values or goods they promote. For instance, a university has certain rights – to enter contracts, to engage in certain forms of speech, and so on – in this derivative sense. These rights are not justified by reference to the university's standing as a moral person or its intrinsic significance, which it arguably does not have, but rather by reference to the functions the university performs in society. In order for a university to fulfil its mission, which is in society's interest, it needs to be able to employ staff, enter contracts, own buildings and equipment, and so on. The university's derivative rights and standing as a legal person enable it to do these things effectively. What I want to suggest is this:

> **A proposal:** (i) Corporate and AI entities may be given derivative rights and legal personhood under certain circumstances, but no non-derivative rights or full-blown moral personhood. (ii) The criterion for deciding whether to grant certain derivative rights (as well as legal personhood) to such entities should be whether this promotes the interests of those of who matter intrinsically, paradigmatically human beings but perhaps also other living creatures.[52]

Under this proposal, it becomes a contingent issue whether the assignment of certain rights and legal personhood to AI systems can be justified. The issue boils down to the question of whether having AI rights and legal personhood helps to promote the interests of those who matter intrinsically, especially human beings. If the answer is yes, then AI rights and legal personhood can be justified; if the answer is no, then not. I suspect that, as long as AI systems have only limited autonomy and limited agential capacities, the answer is no. However, with

[52] The corporate version of this proposal is in line with List and Pettit (2011, ch. 8).

increases in the autonomy and agential capacities of such systems, the answer could become yes, with the resulting rights carefully delimited.[53] I believe that the present proposal, as well as the recognition of the distinction between non-derivative and derivative rights, could take some of the heat out of the debate about AI rights and personhood. The case for or against AI rights and legal personhood is structurally similar to the one for or against corporate rights and legal personhood.

6.3. Could there ever be a case for giving non-derivative rights and a full-blown moral status to group agents or AI systems?

I have emphasized that although group agents and AI systems can qualify as intentional agents and perhaps even as moral agents, their status is very different from that of humans and – I would add – non-human animals. The difference, I have assumed, lies in the fact that human beings and some non-human animals matter intrinsically, while group agents and AI systems do not. Accordingly, only humans and perhaps some other animals can have non-derivative rights and the relevant moral status, while group agents and AI systems can have, at most, derivative rights and a thinner legal status. But even if this differential treatment seems commonsensical, is it philosophically defensible?[54]

If we thought that an entity's moral status was grounded in that entity's agential capacities, then we would have to conclude that similar agential capacities – wherever they occur – imply a similar moral status. And so, if group agents or AI systems were to display the same agential capacities as humans, we would have to extend the same rights and moral status to them. To be sure, the agential capacities of *current* corporate and AI entities fall short of human ones. But this is an empirical fact, which might change with new technological or social developments. The upshot would be that (hypothetical or future) group agents or AI systems with agential capacities similar to those of humans would be entitled to rights and a moral status similar to that of humans.

If we find this conclusion problematic and wish to defend the claim that even group agents or AI systems with very sophisticated capacities should not have the same moral status as humans, we must find some distinguishing feature – beyond agential capacities – that accounts

[53] Solum (1992, pp. 1257–1258) makes a version of this point, noting that if "granting AIs [for example] freedom of speech [had] the best consequences for humans" (perhaps "because this action would promote the production of useful information"), then this would be an "easy justification" for giving certain rights to AI systems, without grounding those rights in the autonomy of the AI speakers themselves.

[54] On this debate in the corporate case, see also Pasternak (2017) and Silver (2018), who argues for extending a certain moral status to corporations. In the AI case, see again Solum (1992).

for the difference in status. The feature would have to be one which human beings unambiguously satisfy, and which group agents and AI systems unambiguously violate. Simply taking membership of the human species as a necessary condition for intrinsic moral significance would be rather stipulative and *ad hoc*, and it would amount to a philosophically unsatisfactory form of "human supremacism", as Will Kymlicka has argued in another context.[55] Among other things, it would run the risk of "throw[ing] animals under the bus".[56] For instance, we may plausibly wish to extend non-derivative rights and an appropriate moral status to the great apes as well. What we need is a criterion that excludes group agents and AI systems from having such rights and status, while including all human beings and also making room for the possibility that some non-human animals could have non-derivative rights as well. My suggestion is the following:

> **A proposal:** A necessary (though perhaps not by itself sufficient) condition for having non-derivative rights and intrinsic moral significance is having phenomenal consciousness or at least having the potential for phenomenal consciousness. To have phenomenal consciousness, in turn, is to be an entity that has subjective experiences: there must be something it is like to be that entity, as Thomas Nagel famously puts it.[57]

This condition is very inclusive. It is satisfied by all human beings,[58] and we can safely assume that it is satisfied by many other sentient animals too, from chimpanzees to cats, dogs, and bats (as Nagel points out), though there could be some further necessary conditions for having non-derivative rights that some or many non-human animals do not satisfy. In any case, the inclusiveness of the proposed condition is desirable. When it comes to deciding which entities are at least candidates for having rights and intrinsic moral significance, we ought to err on the side of inclusiveness rather than exclusiveness. At the same time, as intended, the proposed condition is not satisfied by group agents or by current AI systems. We have very good reasons to think that neither group agents nor current-generation AI systems have anything close to phenomenal consciousness or the potential for it. Satisfying the conditions for intentional agency – even for moral agency – is not the same as having first-person subjective experiences. Intentional agency, and moral agency as I have characterized it here, can be viewed as

---

[55] See Kymlicka (2017).
[56] Ibid.
[57] See Nagel (1974). For earlier discussions of this condition for non-derivative rights in the contexts of artificial intelligence and group agency, see, respectively, Solum (1992) and List (2018).
[58] For instance, babies undeniably have the potential for phenomenal consciousness (whether actualized or not), as do comatose patients, provided there is some chance they will regain consciousness. To make the criterion even more inclusive, we might extend the notion of "potential" to include "past potential" as well.

"functionalist" phenomena – phenomena that lend themselves to a third-personal scientific study – while phenomenal consciousness, by being first-personal, is not like this. Phenomenal consciousness is logically distinct from, and goes beyond, intentional or moral agency.[59]

The main reason for believing that neither group agents nor current AI systems have phenomenal consciousness is that the kinds of physical conditions which, according to recent neuroscience, are associated with phenomenal consciousness and which can be found in the brains of mammals, are not present in corporate institutional structures or in the conventional computer systems underlying existing AI technologies. In particular, we have reason to believe the following:

> **An empirical premise:** Consciousness occurs only in highly integrated information-processing systems with massive internal high-bandwidth feedback mechanisms, of the kind we find in the mammalian cortex. This special kind of informational integration is nowhere to be found in existing corporations or conventional computers.[60]

This is certainly not a conceptual truth, but it is an empirical claim that seems supported by our current best neuroscientific understanding of the correlates of consciousness.

Now comes an important, albeit very hypothetical point: if there ever were corporate entities or AI systems exhibiting the consciousness-supporting physical conditions – such as highly integrated information-processing with high-bandwidth internal feedback, on a par with what goes on in a mammalian cortex – then we would have to conclude that those entities have the potential for phenomenal consciousness. And so, they could become candidates for having intrinsic moral significance and non-derivative rights, at least as far as the proposed necessary condition is concerned. I consider this a very hypothetical scenario, but it is worth reflecting about it for a moment.

In the corporate case, it seems unlikely that we could ever design an organized collective – a corporation or similar – in such a way as to replicate the kind of massively integrated information processing characteristic of a biological cortex.[61] Conceivably – though this is philosophically controversial – the so-called "China brain" thought experiment, described by Ned Block, might illustrate a phenomenally conscious collective.[62] In this thought experiment,

---

[59] See, e.g., Chalmers (1996).
[60] I am here relying on integrated information theory (e.g., Tononi and Koch 2015), though other theories of consciousness (including the neural synchronization theory defended in Crick and Koch 1990) might support similar conclusions. For discussion, see List (2018) and Schwitzgebel (2015).
[61] See List (2018). Cf. Schwitzgebel (2015).
[62] See Block (1980).

each member of a very large population (say, billions) is given the task of simulating the behaviour of one neuron in a biological brain, and the appropriate neural connectivity is then achieved via something like the internet. But existing corporate structures do not remotely resemble anything like this, and so we can set the possibility aside as speculative science fiction.

In the AI case, however, the situation is potentially different. Even though current AI technologies do not appear to instantiate the consciousness-supporting physical conditions, technologies such as biological or biomorphic computing, in which computer systems are built around a neural-network structure, might plausibly support the kind of massively integrated information processing that appears to underpin consciousness in a biological cortex. If this is so, then we cannot rule out the possibility of phenomenally conscious AI from the outset. Indeed, scientific projects such as the EU-funded Brain Simulation Project are attempts to create a reasonably faithful computer simulation of a human brain.[63] If my analysis is right, then those projects and similar developments may have serious ethical implications. Future AI systems with brain-like features might have the potential for phenomenal consciousness, and so they might become candidates for having intrinsic moral significance and some non-derivative rights.[64] In ethical terms, this is completely uncharted territory.

7. The bigger picture

Traditionally, we have tended to assume that human beings are the primary – or even only – powerful agents occupying the social world and that, although non-human animals can be intentional agents too, moral agency is a uniquely human phenomenon. Furthermore, we have tended to assume – especially outside animal-rights or deep-ecology circles – that human beings are the only entities that have intrinsic moral significance and non-derivative rights, so that the class of moral agents and the class of those with intrinsic moral significance do not come apart too significantly.[65] Arguably, the latter class (of beings with intrinsic moral significance) is larger than the former (of moral agents), since there are some humans, such as new-born babies or people with certain impairments, who do not currently meet all the functional conditions for the full exercise of moral agency but who, of course, have rights and entitlements as well as intrinsic moral significance. And many of us would further wish to include some non-human animals in the class of beings with intrinsic moral significance. That said, our established moral codes are relatively anthropocentric, and the philosophical quest to

[63] See <https://www.humanbrainproject.eu/en/brain-simulation/>.
[64] Basl and Schwitzgebel (2019) have independently made a similar point.
[65] On animal rights, see, e.g., Gruen (2017).

identify the conditions for moral agency and the grounds of an entity's moral status has largely boiled down to finding a good justification for the status quo or perhaps some more enlightened version of it.[66]

As my discussion, like earlier discussions of group agency and AI in the literature, should illustrate, the arrival of new complex intentional agents – whether corporate or artificial – raises several challenges for the status quo:

Additional powerful agents: The class of powerful agents occupying the social world and making high-stakes decisions may increasingly include a variety of non-human entities, corporate or artificial.

New responsibility gaps: If we do not carefully implement the kinds of regulatory proposals sketched above – making sure that powerful corporate and artificial agents are designed to satisfy the requirements for fitness to be held responsible – we run the risk that certain high-stakes decisions are made by intentional systems that are unaccountable. This may create new responsibility gaps.

Moral agents without intrinsic moral significance: While traditionally all moral agents have been human and thus entities with intrinsic moral significance, some new moral agents – corporate or artificial – may come into existence which lack such a moral status and matter only derivatively.

Non-human users of morality: While, traditionally, moral principles and theories have always been developed for a human audience – with humans as their target users – we will now need to develop moral principles and theories for some non-human users: corporate entities and AI systems. This raises the question of whether our largely anthropocentric moral codes are fit for purpose or whether the fact that the addressees of morality may include non-human entities changes the way in which moral principles and theories should be formulated, codified, and transmitted.

New entities with intrinsic moral significance: In case phenomenally conscious AI systems ever come into existence, we might need to recognize such entities as candidates for having intrinsic moral significance and certain non-derivative rights – an entirely new situation. We must also ask which norms and standards govern the permissibility of bringing such entities into existence.

[66] On the grounds of moral status, see Jaworska and Tannenbaum (2018).

All this suggests that our moral theories and regulatory frameworks should be "future-proofed": they need to be reassessed with a view to making them adequate even in a world in which some of these challenges materialize. This might require a wider "reflective equilibrium" in our thinking about agency, moral status, and the function of morality for regulating behaviour. Undoubtedly, this paper has flagged more questions than it has answered. But these questions should have a firm place on our agenda for discussion, and my programmatic aim has been to make them more salient: we must give greater attention to the ways in which group agency and artificial intelligence introduce new *loci* of agency into our social world.

References

Basl, J., and E. Schwitzgebel. 2019. AIs should have the same ethical protections as animals. *Aeon*. <https://aeon.co/amp/ideas/ais-should-have-the-same-ethical-protections-as-animals>.

Björnsson, G., and K. Hess. 2017. Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents. *Philosophy and Phenomenological Research* 94(2): 273–298.

Block, N. 1980. Troubles with Functionalism? In N. Block, ed., *Readings in Philosophy of Psychology, Vol. 1*: 268–306. London: Methuen.

Boden, M. A. 2016. *AI: Its Nature and Future*. Oxford: Oxford University Press.

Boella, G., and L. van der Torre. 2007. A Game-Theoretic Approach to Normative Multi-Agent Systems. *Normative Multi-agent Systems, Dagstuhl Seminar Proceedings*. <http://drops.dagstuhl.de/opus/volltexte/2007/937/>.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Braham, M., and M. van Hees. 2011. Responsibility voids. *The Philosophical Quarterly* 61(242): 6–15.

Bratman, M. E. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. 1999. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Bratman, M. E. 2014. *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press.

Broersen, J. 2014. Responsible Intelligent Systems. *Künstliche Intelligenz* 28(3): 209–214.

Chalmers, D. J. 1996. *The Conscious Mind*. New York: Oxford University Press.

Chalmers, D. J. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17(9–10): 7–65.

Chokshi, N. 2018. Robot Conquers One of the Hardest Human Tasks: Assembling Ikea Furniture. *New York Times*, 18 April 2018.

Chopra, A., L. van der Torre, and H. Verhagen. 2018. *Handbook of Normative Multiagent Systems*. London: College Publications.

Collins, S. 2017. Filling Collective Duty Gaps. *Journal of Philosophy* 114(11): 573–591.

Collins, S. 2018. Collective Responsibility Gaps. *Journal of Business Ethics*. Online.

Colvin, E. 1995. Corporate Personality and Criminal Liability. *Criminal Law Forum* 6: 3–44.

Copp, D. 2006. On the Agency of Certain Collective Entities: An Argument from "Normative Autonomy". *Midwest Studies in Philosophy* 30(1): 194–221.

Crick, F., and C. Koch. 1990. Towards a Neurobiological Theory of Consciousness. *Seminars in the Neurosciences* 2: 263–275.

Delcker, J. 2018. Europe divided over robot "personhood". *Politico*. <https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/>.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. 2009. Intentional Systems Theory. In A. Beckermann, B. P. McLaughlin, and S. Walter, eds., *The Oxford Handbook of Philosophy of Mind*: 339–350. Oxford: Oxford University Press.

Duijf, H. 2018. Responsibility Voids and Cooperation. *Philosophy of the Social Sciences*. Online.

Dretske, F. I. 1999. Machines, Plants and Animals: The Origins of Agency. *Erkenntnis* 51(1): 523–535.

Erskine, T. 2001. Assigning Responsibilities to Institutional Moral Agents: The Case of States and Quasi-States. *Ethics & International Affairs* 15(2): 67–85.

Fine, A. 1984. The Natural Ontological Attitude. In J. Leplin, ed., *Philosophy of Science*: 261–277. Berkeley: University of California Press.

Fisher, M., C. List, M. Slavkovik, and A. Winfield. 2016. Engineering Moral Machines. *Informatik Spektrum* 39(6): 467–472.

Fossa, F. 2018. Artificial moral agents: moral mentors or sensible tools? *Ethics and Information Technology*. Online.

French, P. A. 1984. *Collective and Corporate Responsibility*. New York: Columbia University Press.

Gilbert, M. 1989. *On Social Facts*. New York: Routledge.

Gruen, L. 2017. The Moral Status of Animals. In E. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition) <https://plato.stanford.edu/archives/fall2017/entries/moral-animal/>.

Jaworska, A., and J. Tannenbaum. 2018. The Grounds of Moral Status. In E. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition) <https://plato.stanford.edu/archives/spr2018/entries/grounds-moral-status/>.

Kymlicka, W. 2017. Human rights without human supremacism. *Canadian Journal of Philosophy*. Online.

Laukyte, M. 2014. Artificial Agents: Some Consequences of a Few Capacities. In J. Seibt et al., eds., *Sociable Robots and the Future of Social Relations*: 115–122. IOS Press.

Laukyte, M. 2017. Artificial agents among us: Should we recognize them as agents proper? *Ethics and Information Technology* 19(1): 1–17.

List, C. 2018. What is it like to be a group agent? *Noûs* 52(2): 295–319.

List, C. 2019. Group Responsibility. Manuscript, London School of Economics.

List, C., and P. Pettit. 2006. Group Agency and Supervenience. *Southern Journal of Philosophy* 44(S1): 85–105.

List, C., and P. Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Matthias, A. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

Minsky, M. 1986. *The Society of Mind*. New York: Simon and Schuster.

Nagel, T. 1974. What Is It Like to Be a Bat? *Philosophical Review* 83(4): 435–450.

Olson, M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1977. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Parikh, R. 2002. Social Software. *Synthese* 132(3): 187–211.

Pasternak, A. 2017. From Corporate Moral Agency to Corporate Moral Rights. *The Law & Ethics of Human Rights* 11(1): 135–159.

Pettit, P. 2001. *A Theory of Freedom: From the Psychology to the Politics of Agency*. Cambridge and New York: Polity and Oxford University Press.

Pettit, P. 2003. Groups with Minds of their Own. In F. Schmitt, ed., *Socializing Metaphysics*: 167–193. New York: Rowan and Littlefield.

Pettit, P. 2007. Responsibility Incorporated. *Ethics* 117(2): 171–201.

Pinker, S. 2018. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. New York: Penguin.

Rovane, C. 1997. *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton: Princeton University Press.

Russell, S. J., and P. Norvig. 2009. *Artificial Intelligence: A Modern Approach*. 3rd edition. Upper Saddle River, NJ: Prentice Hall Press.

Schwitzgebel, E. 2015. If Materialism is True, the United States is Probably Conscious. *Philosophical Studies* 172(7): 1697–1721.

Silver, K. 2018. Can a Corporation be Worthy of Moral Consideration? *Journal of Business Ethics*. Online.

Smiley, M. 2017. Collective Responsibility. In E. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition) <https://plato.stanford.edu/archives/sum2017/entries/collective-responsibility/>.

Solum, L. B. 1992. Legal Personhood for Artificial Intelligences. *North Carolina Law Review* 70(4): 1231–1287.

Sparrow, R. 2007. Killer Robots. *Journal of Applied Philosophy* 24(1): 62–77.

Tollefsen, D. P. 2002. Collective Intentionality and the Social Sciences. *Philosophy of the Social Sciences* 32(1): 25–50.

Tollefsen, D. P. 2015. *Groups as Agents*. Cambridge: Polity Press.

Tononi, G., and C. Koch. 2015. Consciousness: Here, There and Everywhere? *Philosophical Transactions of the Royal Society* B 370.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460.

Turner, J. 2018. *Robot Rules: Regulating Artificial Intelligence*. Heidelberg: Springer.

Tuomela, R. 2007. *The Philosophy of Sociality: The Shared Point of View*. New York: Oxford University Press.

Tuomela, R. 2013. *Social Ontology: Collective Intentionality and Group Agents*. Oxford: Oxford University Press.

Valentini, L. 2018. Why the notion of moral claim rights is unhelpful. Manuscript, London School of Economics.

Winfield, A. F. T. 2012. *Robotics: A Very Short Introduction*. Oxford: Oxford University Press.

Winfield, A. F. T., and M. Jirotka. 2017. The Case for an Ethical Black Box. In Y. Gao, S. Fallah, Y. Jin, and C. Lekakou, eds., *Towards Autonomous Robotic Systems, TAROS 2017*. *Lecture Notes in Computer Science*. Springer.